Statistics for Data Analysis

Beginner to Advanced

1. What is Statistics?

Statistics is the science of collecting, analyzing, presenting, and interpreting data. It allows us to make sense of the vast amounts of information we encounter in various fields.

Data: We collect, analyze, and summarize these facts and figures. Data can be classified as quantitative or qualitative.

Variables: Characteristics like age, gender, marital status, and annual income are called variables. Each individual has associated data values for these variables.

Quantitative vs. Qualitative:
Quantitative variables (like age and income) have numerical values.
Qualitative variables (like gender and marital status) provide labels or categories.

Sample Surveys and Experimental Studies:
Sample survey methods collect data from observational studies.
Experimental design methods collect data from experimental studies.

In summary, Statistics helps us turn raw data into meaningful information, guiding decision-making and problem-solving.

2. Types of Statistics?

Descriptive Statistics:
Descriptive statistics involves summarizing and organizing data to gain insights. It's like taking a snapshot of the data.

Purpose: Descriptive statistics helps us understand the main features of a dataset, such as central tendency (mean, median, mode), variability (range, variance, standard deviation), and distribution.

Inferential Statistics:
Inferential statistics goes beyond describing data; it allows us to make predictions and draw conclusions about a larger population based on a sample.

Purpose: Inferential statistics helps us infer properties of a population from a smaller subset (sample) of that population.

3. Population and Samples?

Population:

A population refers to the entire group that you want to draw conclusions about. It encompasses all the individuals, objects, events, or elements relevant to your study.

Examples:

In a study about job advertisements for IT positions in the Netherlands, the population would include all such advertisements available on a specific date.

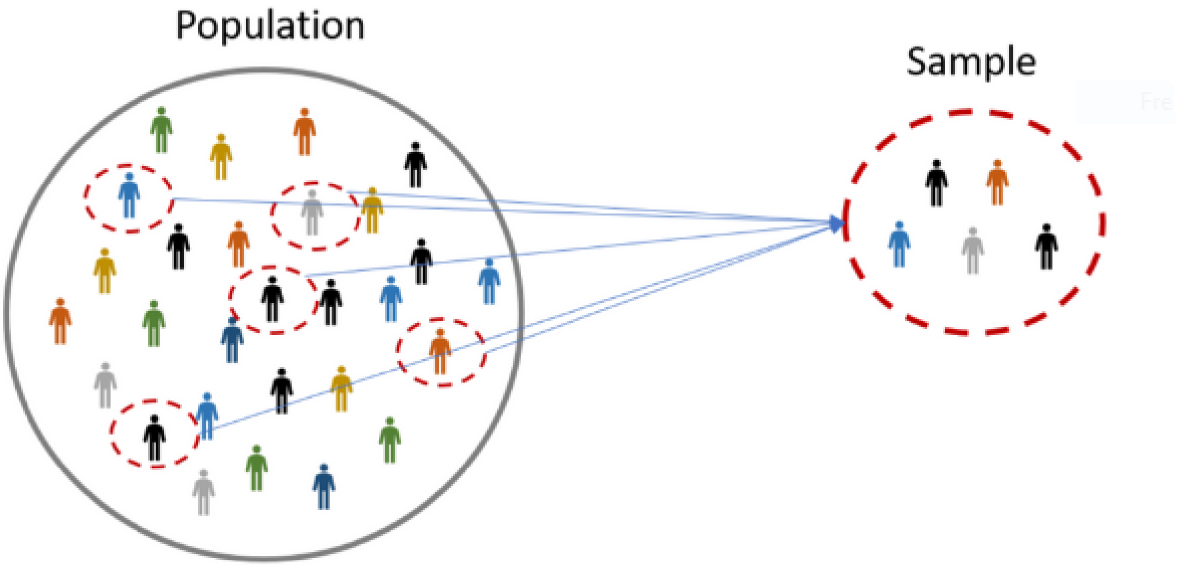
Sample:

A sample is a specific subset of the population from which you collect data. It's practically impossible to gather information from every individual in a large or dispersed population.

We use samples to make inferences about the entire population.

Examples:

The Census, conducted every decade, aims to count every person living in the country. However, due to challenges in reaching marginalized and low-income groups, the actual count remains incomplete and biased. In such cases, sampling helps make more precise inferences.



4. Central measure of tendency?

Mean (Average):

The mean is calculated by adding up all the values in the dataset and then dividing by the total number of values.

It represents the central value around which the data points tend to cluster.

Formula: $$\text{Mean} = \frac{\sum \text{values}}{\text{total number of values}}$$

Example: If we have exam scores of 80, 85, 90, and 95, the mean score would 480+85+90+95 / 4 =87.5

Median:

The median is the middle value when the dataset is arranged in ascending or descending order.

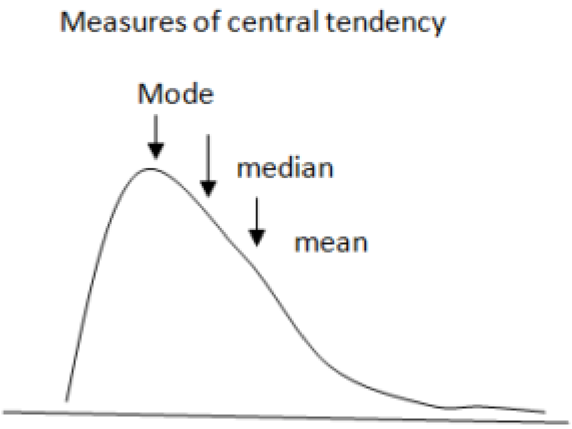
If there's an even number of values, the median is the average of the two middle values. It's less sensitive to extreme values (outliers) than the mean.

Example: For the dataset {10, 20, 30, 40, 50}, the median is 30.

Mode:

The mode is the most frequent value in the dataset.
A dataset can have no mode, one mode, or multiple modes.

Example: In a survey, if the responses for political affiliation are {Conservative, Moderate, Liberal, Moderate}, and "Moderate" appears most frequently, it's the mode.



5. Dispersion?

Dispersion refers to the degree of variability or spread in a dataset. It tells us how the data points are distributed around a central value (such as the mean, median, or mode).

Understanding dispersion is crucial because it provides insights into the variability within the data.

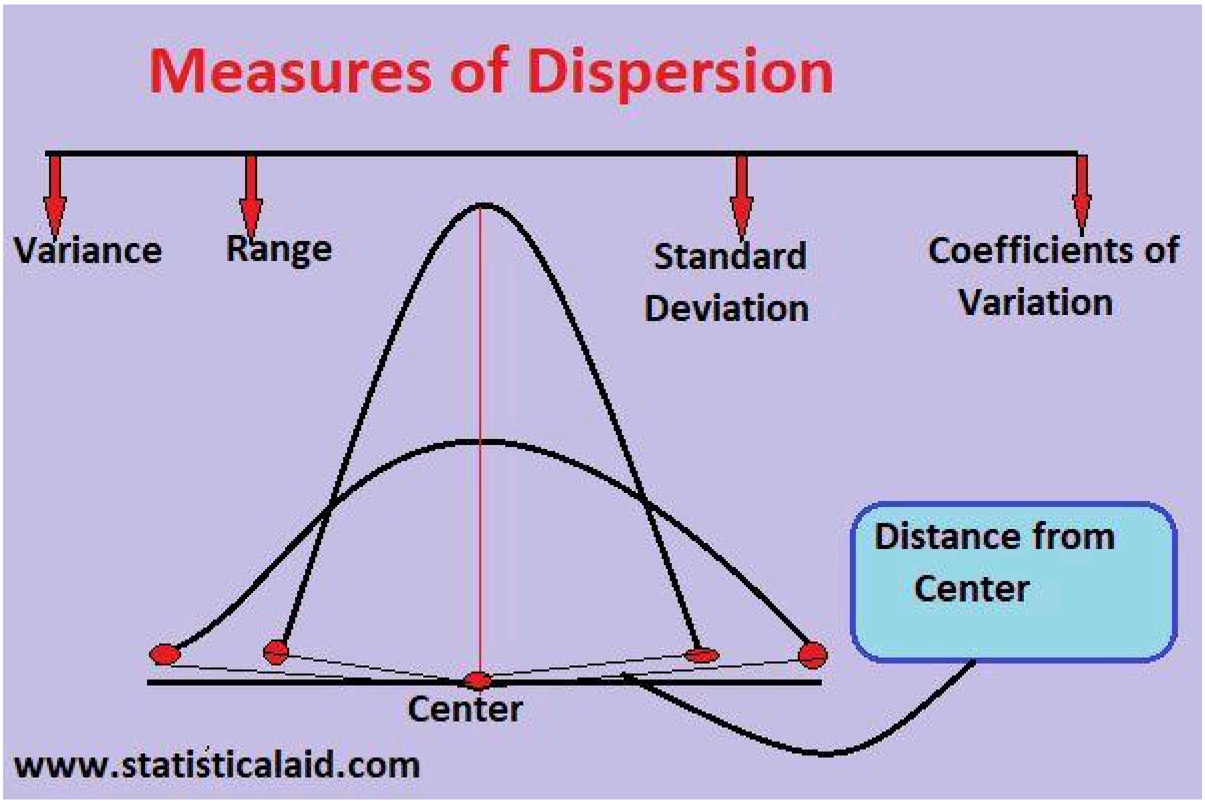
Just like central tendency measures summarize the center of the data, dispersion measures summarize its spread.

Example: Imagine a dataset of exam scores for two classes:

Class A: {80, 85, 90, 95, 100}
Class B: {60, 70, 80, 90, 100}

Both classes have the same mean (average) score of 90, but Class B has greater dispersion because its scores are more spread out.

Measures of Dispersion:
These measures quantify how data points far from the central value. Here are some common ones:



Range: The difference between the maximum and minimum values in the dataset.

Variance: The average of the squared differences between each data point and the mean.

Standard Deviation: The square root of the variance. It indicates the typical deviation from the mean.

Interquartile Range (IQR): The range of the middle 50% of the data (between the 25th and 75th percentiles).

Coefficient of Variation (CV): The ratio of the standard deviation to the mean (expressed as a percentage).

Mean Absolute Deviation (MAD): The average of the absolute differences between each data point and the mean.

Why Measure of Dispersion essential to understand?

Dispersion measures help us:

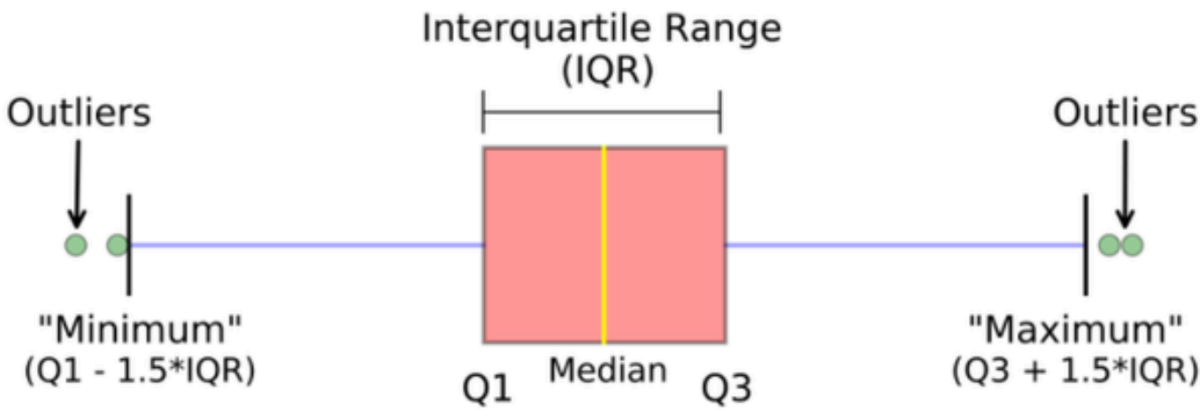
1. Assess the variability within a dataset.
2. Identify outliers (extreme values).
3. Make informed decisions based on the spread of data.

Remember, while central tendency measures give us a snapshot of the center, dispersion measures reveal how the data is scattered. Both aspects are essential for a comprehensive understanding of any dataset.

6. Quartile:

Quartiles divide an ordered dataset into four equal parts. They help us understand the distribution of data by identifying key points and help to detect the outliers in the data.

Q1 (First Quartile): Separates the lowest 25% of values from the rest. It’s equivalent to the 25th percentile.
Q2 (Second Quartile): This is the median, dividing the data into the bottom and top halves. It’s equivalent to the 50th percentile.
Q3 (Third Quartile): Separates the lowest 75% from the highest 25%. It’s equivalent to the 75th percentile.



7. What is Distribution?

A distribution shows the possible values of a variable and how often they occur. Think of it as a way to visualize the likelihood of different outcomes.

Types of Distributions:

There are different types of distributions. Defining some of them.

1. Discrete Distributions:
These apply to variables with countable outcomes (e.g., whole numbers).

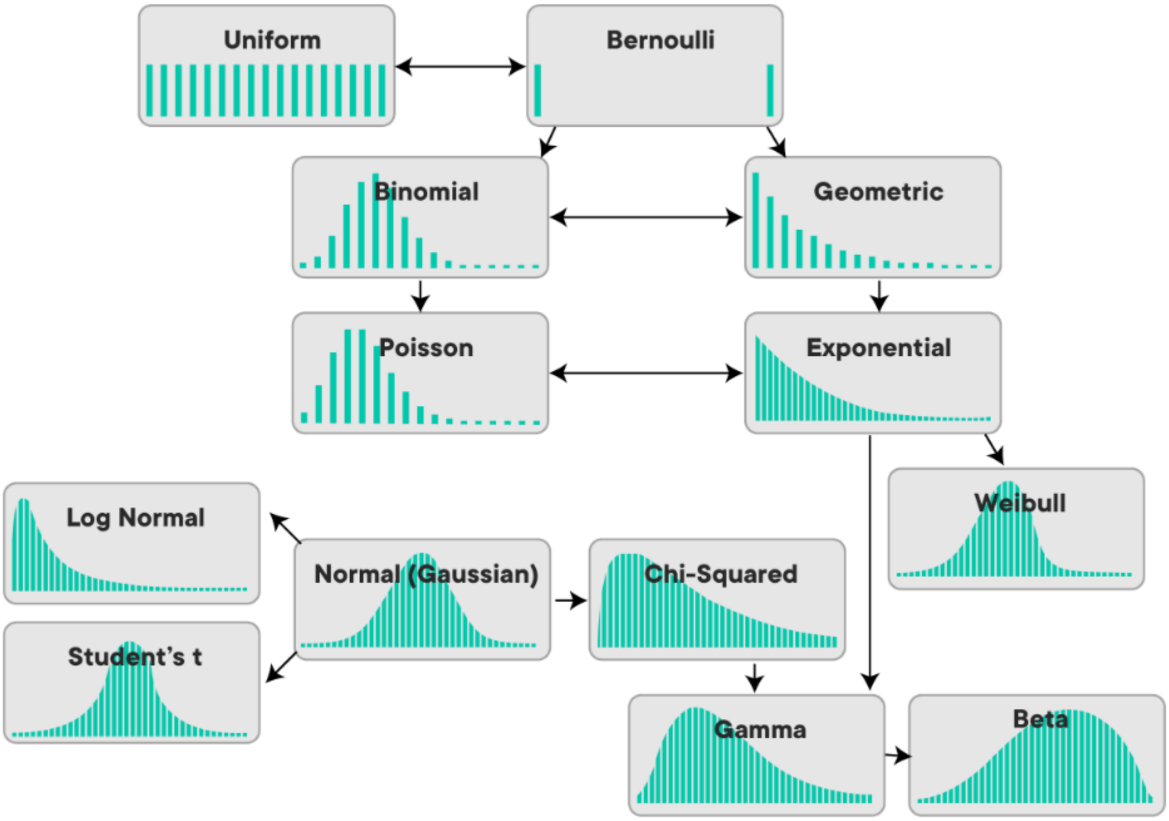
Examples:

- **Binomial Distribution:** Models the number of successes in a fixed number of independent trials (like Tossing a coin).
- **Poisson Distribution:** Describes rare events occurring over a fixed interval (e.g. number of emails received per hour).

2. Continuous Distributions:
These apply to variables with infinite possible outcomes (e.g., real numbers).

Examples:

- **Normal (Gaussian) Distribution:** Often seen in natural phenomena (a bell-shaped curve).
- **Uniform Distribution:** All values have equal probability (like rolling a fair die).



Distributions help us:

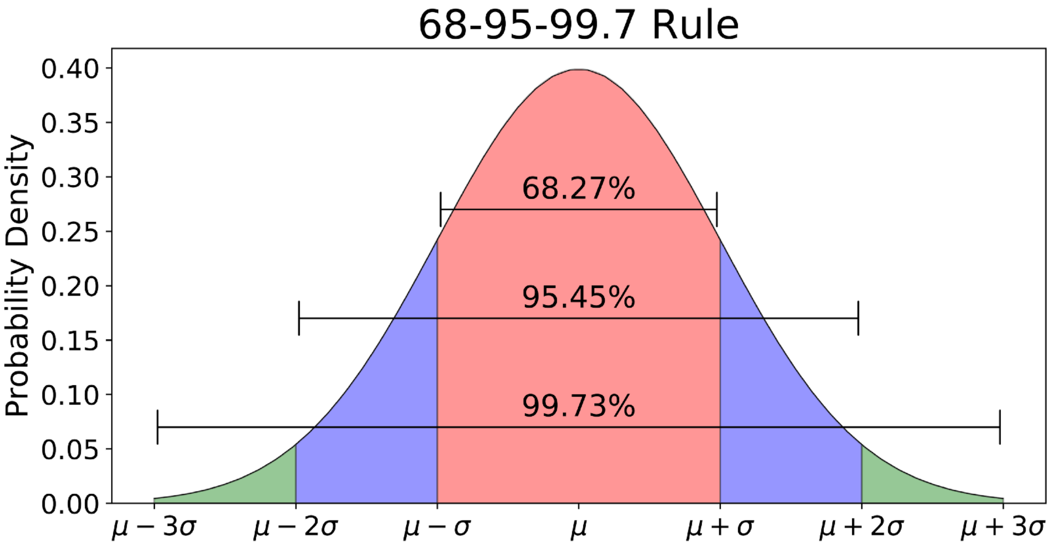
1. Understand the central tendency (mean, median) and spread (variance, standard deviation) of data.
2. Make predictions and estimate probabilities.
3. Model real-world phenomena (from egg sizes to stock prices).

Example:
Imagine an egg farmer weighing 100 random eggs. She creates a histogram showing the distribution of egg weights.

From this distribution, she can estimate the probability of different egg sizes.

8. What is Probability?

Probability is a measure of the likelihood of an event occurring. It ranges from 0 (indicating an impossible event) to 1 (representing a certain event). In other words, probability helps us predict how likely something is to happen.



Here’s the basic formula for calculating probability:

$$\text{Probability of an event (P)} = \frac{\text{Total number of possible outcomes}}{\text{Number of favorable outcomes}}$$

For example:

If there are 6 pillows on a bed (3 red, 2 yellow, and 1 blue), the probability of picking a yellow pillow is 1/3.

9. What is Hypothesis Testing?

Hypothesis testing is a formal procedure used in statistics to investigate our ideas about the world. It helps us evaluate specific predictions (called hypotheses) that arise from theories.

Here are the key steps involved in hypothesis testing:
State of Hypotheses:

Null Hypothesis (H_0): This predicts no relationship between the variables you're interested in. It's often denoted as H_0.
Alternate Hypothesis (H_a or H_1): This predicts a specific relationship between the variables. It's your initial hypothesis.

Example:

Suppose you want to test whether men are, on average, taller than women. Your hypotheses would be:

H_0: Men are, on average, not taller than women.
H_a: Men are, on average, taller than women.

Collect Data:

Gather data in a way that is designed to test your hypothesis.

Representative sampling is crucial for valid results.
For example, if you're comparing average heights between men and women, ensure your sample includes both genders and covers various socio-economic classes.

Perform a Statistical Test:

Choose an appropriate statistical test based on your data and research question. These tests compare within-group variance (spread of data within a category) to between-group variance (differences between categories).

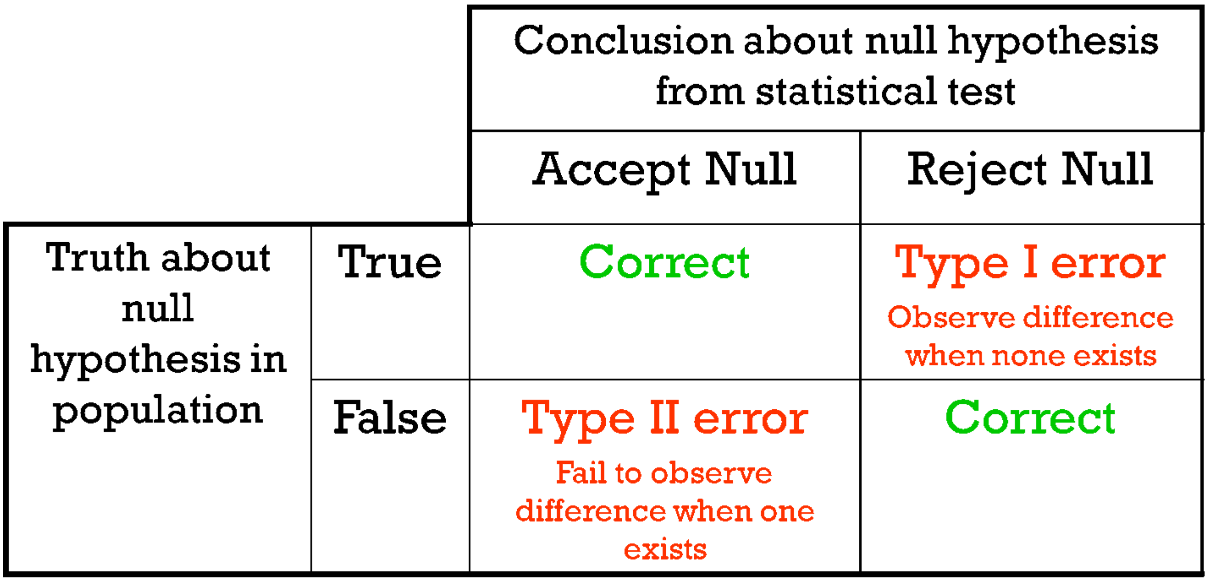
Decide Whether to Reject or Fail to Reject the Null Hypothesis:

Based on the test results, you'll either:

Reject H_0: If the evidence strongly supports the alternate hypothesis.
Fail to Reject H_0: If there isn't enough evidence to support the alternate hypothesis.

Present Your Findings:

Communicate the results in your research report or discussion section. Be clear about which hypothesis you're supporting based on the data.



		Conclusion about null hypothesis from statistical test	
		Accept Null	Reject Null
Truth about null hypothesis in population	True	Correct	Type I error Observe difference when none exists
	False	Type II error Fail to observe difference when one exists	Correct

Explanation:

Type I Error (False Positive):

Occurs when we incorrectly reject the null hypothesis (H_0) when it is actually true. In other words, we conclude there's an effect or difference when there isn't one.

Type II Error (False Negative):
Occurs when we fail to reject the null hypothesis (H_0) when it is actually false. In other words, we miss a real effect or difference.

P-Value:

The p-value measures the strength of evidence against the null hypothesis. It represents the probability of observing the data (or more extreme data) if the null hypothesis were true.

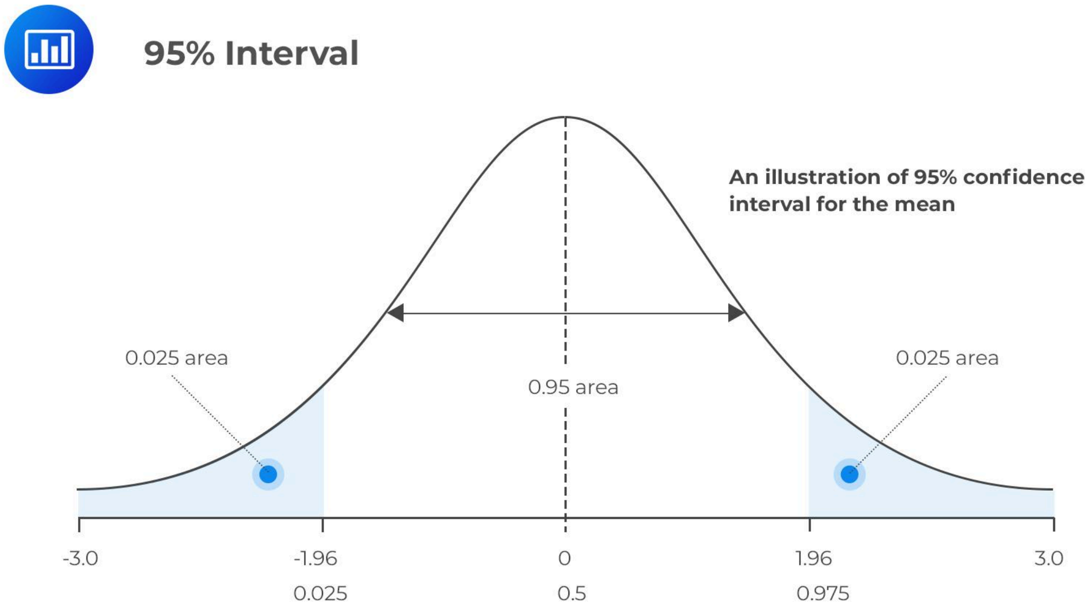
If p-value < α (significance level), we reject H_0.

Example: A p-value of 0.035 means there's a 3.5% chance of observing the data if H_0 is true.

Confidence Interval:

A confidence interval (CI) provides a range of values within which we believe the true population parameter lies. It quantifies our uncertainty about the estimate.

Example: A 95% CI for the average height of students might be (160 cm, 170 cm).



Z-Test and T-Test:

Z-Test:
Used when we know the population standard deviation (σ). Compares a sample mean to a known population mean.

Example: Testing if a new drug's effectiveness differs from the standard treatment.

T-Test:
Used when we don't know the population standard deviation (use sample standard deviation, s). Compares means of two groups (independent samples) or before/after treatment (paired samples).

Example: Comparing exam scores between two teaching methods.

Scenario: Analyzing Customer Satisfaction at an E-Commerce Company

Background:
An e-commerce company wants to improve customer satisfaction. Collect data on customer reviews, ratings, and purchase behavior.

Objective:
1. Understand factors affecting customer satisfaction.
2. Identify areas for improvement.

Data Collection:
The company gathers data from:
1. Customer reviews (textual feedback).
2. Ratings (1 to 5 stars).
3. Purchase history (products bought, order frequency).

Exploratory Data Analysis (EDA):	**Results of EDA:**
1. Descriptive Statistics: 1. Calculate mean, median, and mode of ratings. 2. Visualize the distribution of ratings (histogram).	**Descriptive Stats:** Average rating: 4.2 stars.
2. Word Clouds: Create word clouds from customer reviews to identify common themes (positive/negative).	**2. Word Clouds:** Most common words in reviews: “fast,” “quality,” “service.”
3. Correlation Analysis: Check if higher ratings correlate with more frequent purchases.	**Correlation:** Positive correlation between ratings and purchase frequency.
4. Hypothesis Testing: 1. Hypothesis: Higher ratings lead to increased repeat purchases. 2. Perform a t-test comparing average ratings for repeat customers vs. one-time customers.	**Hypothesis Test:** Reject null hypothesis ($p < 0.05$): Higher ratings are associated with more repeat purchases.

Recommendations:

1. Improve product quality and delivery speed.
2. Address specific issues mentioned in negative reviews.
3. Implement loyalty programs to encourage repeat purchases.

Conclusion:

Data analysis reveals actionable insights for enhancing customer satisfaction. The company can now focus on targeted improvements.